Energy Consumption Forecasting

Balaji Vijayaraj
Data Science
Dalarna University
Borlänge, Sweden
v22balvi@du.se

Abstract— **Accurate forecasting of energy consumption plays a pivotal role in optimizing resource planning, decision-making, and operational efficiency in the energy industry. This project aims to develop a predictive model using the XGBoost and SARIMAX algorithms to accurately forecast energy consumption for a given time period. Leveraging time series analysis and machine learning techniques, the model captures the complex and dynamic nature of energy systems, enabling more reliable predictions. The project identifies areas where the model exhibits deviations or inaccuracies in its predictions. The results showcase the effectiveness of the proposed approach, with promising performance metrics on both training and test sets. Furthermore, the project explores the top errors and provides insights into areas where improvements can be made. By addressing these challenges and leveraging advanced techniques, this project contributes to cost reduction, sustainability, and improved energy management strategies in the energy industry. The findings offer valuable insights for stakeholders in optimizing energy supply chains and implementing demand response programs.**

Keywords: energy consumption, time series analysis, features, XGBoost, SARIMAX, root mean squared error (RMSE), mean absolute error (MAE).

I. INTRODUCTION

Energy consumption forecasting plays a crucial role in optimizing the energy supply chain, reducing costs, and increasing operational efficiency in the energy industry. Accurate predictions of energy consumption enable stakeholders to make informed decisions and implement effective strategies for energy management. This project aims to address the challenge of energy consumption forecasting by developing a predictive model using the XGBoost and SARIMAX algorithms. The model leverages time series analysis and machine learning techniques to provide accurate forecasts of energy consumption for a given time period.

The importance of accurate energy consumption forecasting is highlighted in various studies [1][2][3]. These studies emphasize the significance of forecasting strategies and models in optimizing energy usage and improving resource planning. The XGBoost and SARIMAX algorithm has demonstrated their effectiveness in various domains, including energy forecasting [4][5]. By utilizing these algorithms, the project aims to leverage the strengths of XGBoost and SARIMAX in capturing complex patterns and dependencies in energy consumption data.

The aim of this research is to compare the SARIMAX and XGBoost in forecasting energy consumption. Additionally, the project focuses on identifying areas where the predictive models exhibit deviations or inaccuracies in their forecasts. Understanding these limitations helps refine the model and improve the accuracy of energy consumption predictions.

This project contributes to the field of energy consumption prediction by leveraging advanced machine learning techniques to develop a robust and accurate forecasting model. The findings of this research can support decision-making processes, resource planning, and operational efficiency in the energy industry. Furthermore, the insights gained from this project can inform the development of more sophisticated energy management strategies, leading to cost reduction and improved sustainability.

II. LITERATURE REVIEW

Energy consumption forecasting using machine learning techniques has gained significant attention in recent years due to its potential to optimize energy supply chains, reduce costs, and improve overall efficiency. This literature review aims to explore relevant studies and research articles that have investigated energy consumption forecasting, with a particular focus on the application of XGBoost, a popular gradient boosting algorithm and SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables)

Hadri et al. [1] conducted a study on the performance evaluation of forecasting strategies for electricity consumption in buildings. They compared different forecasting models and techniques, including XGBoost, and assessed their accuracy in predicting energy consumption. The results showed that XGBoost outperformed other models in terms of accuracy and reliability.

Bassi et al. [2] compared gradient boosting models for building energy consumption forecasting. They evaluated the performance of different gradient boosting algorithms, including XGBoost, on real-world energy consumption data. The study demonstrated that XGBoost produced accurate and robust forecasts, making it a suitable choice for energy consumption prediction in buildings.

Shin and Woo [3] focused on energy consumption forecasting in Korea using machine learning algorithms. They explored various machine learning models, including XGBoost, to predict electricity consumption in different regions of Korea. The results highlighted the effectiveness of XGBoost in accurately forecasting energy consumption, enabling better energy management and resource allocation.

Abbasi et al. [4] presented a study on short-term load forecasting using XGBoost. They applied XGBoost to predict electricity load demand based on historical data. The study demonstrated that XGBoost achieved high forecasting accuracy and outperformed other traditional forecasting techniques, making it a valuable tool for load forecasting applications.

El Houda et al.[5] conducted a time series analysis of household electric consumption using the XGBoost model. They applied XGBoost to household energy consumption data and analyzed its performance in capturing temporal patterns and predicting future consumption. The results indicated that XGBoost effectively captured complex patterns and exhibited strong predictive capabilities for household energy consumption forecasting.

Amitasree et al.[6] investigated electricity consumption forecasting using machine learning techniques. They explored the application of various machine learning algorithms, including XGBoost, for predicting electricity consumption patterns. The study highlighted the accuracy and efficiency of XGBoost in forecasting electricity consumption, offering insights for efficient energy management.

Elamin et al. [7] focuses on utilizing SARIMAX models with interaction effects to model and forecast hourly electricity demand. The authors emphasize the significance of considering the interactions between exogenous variables, such as weather conditions and holidays, and the time series components in accurately capturing electricity consumption patterns. By incorporating interaction terms in the SARIMAX model, they aim to capture non-linear relationships and dependencies between different factors, thereby improving the forecasting performance. The paper likely discusses the methodology for implementing SARIMAX models with interaction effects, including data preprocessing, model estimation, and validation procedures. Empirical results are also expected, including evaluations of forecasting accuracy and potential comparisons with alternative models or techniques. Overall, Elamin and Fukushige's study contributes to the literature on electricity demand forecasting by highlighting the effectiveness of SARIMAX models with interaction effects in capturing the complexities of hourly demand patterns.

The reviewed studies demonstrate the effectiveness of XGBoost and SARIMAX in energy consumption forecasting. Both algorithms have shown superior performance in accurately predicting energy consumption, enabling better decision-making, and cost reduction in the energy industry. These findings provide a strong foundation for utilizing XGBoost and SARIMAX in developing a predictive model for energy consumption forecasting.

III. METHOD DESCRIPTION

A. The Dataset

The database used in the project is sourced from American Electric Power (AEP) and contains information about estimated energy consumption in Megawatts (MW) [8]. The dataset consists of two columns: "Datetime" and "PJMW_MW".

Datetime: This column represents the timestamp or date and time at which the energy consumption data was recorded. It provides information about the temporal aspect of energy consumption, allowing for time series analysis and forecasting.

PJMW_MW: This column represents the estimated energy consumption in Megawatts (MW). It provides the numerical value of energy consumption at each timestamp. This variable is the main target variable in the project, as the goal is to forecast energy consumption based on historical data.

The dataset is cleaned and prepared for processing through a methodology that involves the creation of time series features. To accomplish this, a function called create_features is implemented. The function copies the original DataFrame to avoid altering the original data and then extracts various time-based features from the time series index. These features include 'hour', 'dayofweek', 'quarter', 'month', 'year', 'dayofyear', 'dayofmonth', and 'weekofyear'. By applying this function to the DataFrame, the time-based features are added as additional columns, providing additional contextual information related to the time component of the data. This modified data frame, now enriched with time series features, is then available for further analysis and modelling tasks, specifically for energy consumption forecasting.

The methodology for removing outliers from the dataset involved loading the data, setting the 'Datetime' column as

the index, and converting it to datetime format. Outliers were detected by calculating the average and standard deviation of the 'PJMW_MW' column. A histogram plot was generated to visualize the data distribution, and lines representing the 3rd standard deviation were marked on the histogram. Outliers were filtered by selecting values less than 2000 or greater than 10000 in the 'PJMW_MW' column, which were plotted separately. The dataset was then updated to remove outliers by keeping only values greater than 2000. Finally, the filtered data was plotted to visualize energy consumption without outliers.

B. Data Mining Method

In this work, the data mining models used to solve the energy consumption forecasting problem are XGBoost and SARIMAX. The XGBoost model is tuned by setting various parameters such as the base score, booster, number of estimators, early stopping rounds, objective function, maximum depth, and learning rate. The SARIMAX model is configured by setting specific Order and seasonal order parameters to achieve the desired performance.

The data is invoked by the model by splitting it into training and test sets using a TimeSeriesSplit object with 5 splits. The dataset is sorted by the time series index to ensure chronological order, which is necessary for TimeSeriesSplit which is depicted in Figure I. The features used for training the model include 'hour', 'dayofweek', 'weekofyear', 'month', and 'year', while the target variable is 'PJMW_MW', representing estimated energy consumption in Megawatts.

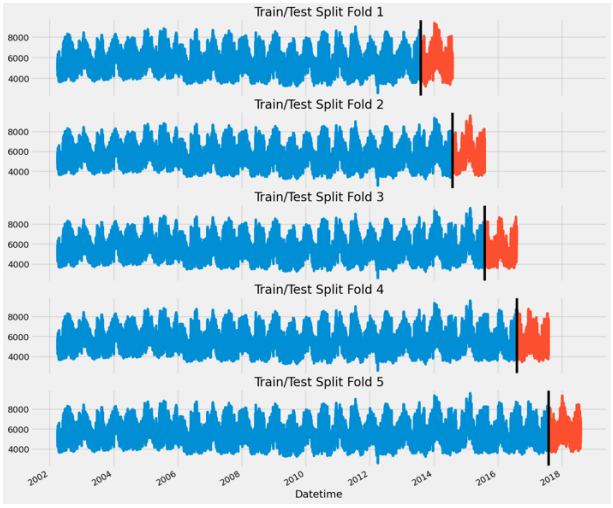

Figure 1: Train/Test Split for Time Series Data

Cross-validation is performed by iterating over the time series splits obtained from TimeSeriesSplit as shown in Figure 1. For each split, the XGBoost and SARIMAX model is fitted on the training data and evaluated on both the training and test sets using the provided evaluation set. The model predicts energy consumption on the test set, and evaluation metrics such as root mean squared error (RMSE), mean absolute error (MAE), and R2 score is calculated to assess the model's performance.

The RMSE and MAE scores on the test set are calculated and printed, providing an assessment of the model's accuracy in predicting energy consumption. Finally, the top 50 errors, measured by the absolute difference between the predicted and actual energy consumption, are identified and displayed for further analysis.

In the retraining process, the model was retrained using the entire dataset to enhance its performance. The features used for training remained the same, including 'hour', 'dayofweek', 'weekofyear', 'month', and 'year', while the target variable was 'PJMW_MW'. The model was initialized with specific parameters and fitted to the full dataset. To generate future predictions, a new DataFrame called 'future_df' was created with future timestamps using the pd.date_range function. This DataFrame was combined with the original DataFrame, and the resulting DataFrame was processed to extract time-based features. The 'future_features' DataFrame, consisting of only the future timestamps, was used to make predictions using the retrained model. The predicted values were added as a new column, and a line plot was generated to visualize the future predictions for energy consumption. By following these steps, the model was retrained and utilized to generate insights into future energy consumption patterns.

IV. Results and Analysis

The XGBoost model achieved an RMSE score of 685.72 and an MAE score of 508.95 on unseen data, indicating its performance on generalization. These scores suggest that the model may struggle to accurately predict energy consumption patterns on new data, as the values are higher than the corresponding scores on the training set. This suggests potential limitations in capturing the underlying patterns and dynamics of the data.

In contrast, the SARIMAX model yielded an RMSE of 149.11 and an MAE of 83.65. These scores indicate that the SARIMAX model outperformed the XGBoost model in terms of both RMSE and MAE. The lower values of RMSE and MAE suggest that the SARIMAX model better captures the energy consumption patterns and provides more accurate predictions compared to the XGBoost model.

Based on this comparison, it is seen that the ARIMA model shows better performance in terms of forecasting accuracy for the given energy consumption dataset. However, it is important to note that the suitability of a model can vary depending on the specific dataset, the patterns within it, and the forecasting requirements.

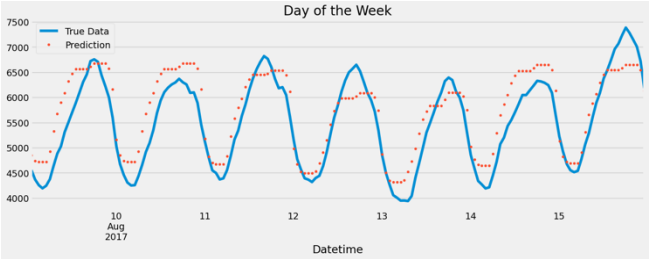

Figure 2 : True Data vs. Predictions of Energy Consumption with XGBoost

Figure 2 focuses on a specific time period, specifically from August 9th, 2017 to August 16th, 2017. It displays the actual energy consumption values (true data) and the corresponding predicted values during this period.

The blue line represents the actual energy consumption data, showing the variation in energy consumption over the selected time range. The orange dots represent the predicted values of energy consumption generated by the model. By plotting the true data and predictions together, we can visually compare how well the model's predictions align with the actual energy consumption during this specific week.

The findings indicate that the XGBoost model performs well in capturing the cyclical patterns of energy consumption. It accurately predicts the general trends and variations in energy usage over different time periods. However, the model tends to be conservative when it comes to forecasting peak consumption periods. It may underestimate the magnitude of energy consumption during these high-demand periods.

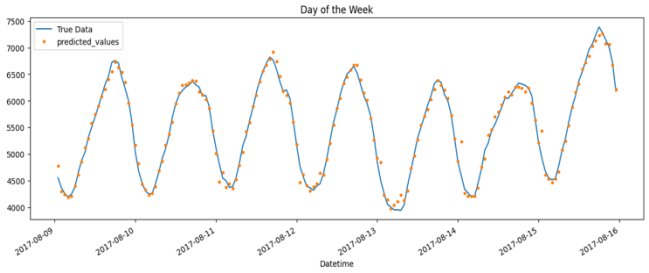

Figure 3: True Data vs. Predictions of Energy Consumption with SARIMAX

Figure 3 presents the comparison between the predicted values and true values using the SARIMAX model for the same time period. The figure clearly illustrates that the predicted values closely align with the true values, indicating the model's ability to accurately capture the patterns and trends in the data.

Figures 2 and 3 suggest that the SARIMAX model excels in capturing both the cyclical patterns and peak consumption periods of energy usage accurately. It effectively predicts the general trends and variations in energy consumption over different time periods while also providing reliable forecasts for high-demand periods. Unlike the XGBoost model, the SARIMAX model exhibits a better ability to capture the magnitude of energy consumption during peak periods, resulting in more accurate forecasts. This capability of accurately capturing both cyclical patterns and peak consumption periods makes the SARIMAX model a suitable choice for energy consumption forecasting tasks.

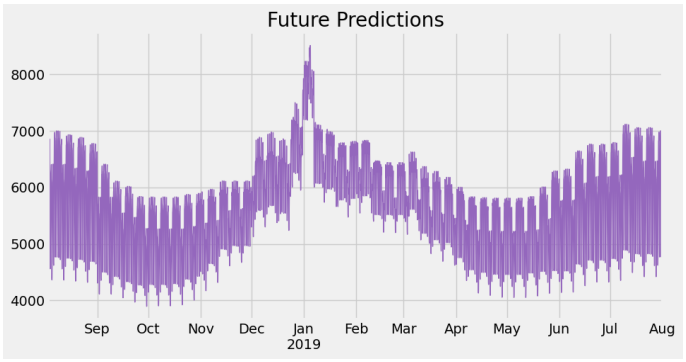

Figure 4 : Future Energy Consumption Predictions

Figure 4 provides a visual representation of the predicted values for future energy consumption based on the trained model. It offers insights into the anticipated energy demand patterns in the future. From this plot, stakeholders and decision-makers can gain a clear understanding of the expected energy consumption trends, enabling them to make informed decisions regarding resource allocation, energy generation, and distribution. However, it is important to recognize that these predictions are based on historical patterns and real-world factors may influence actual energy consumption. Therefore, the chart should be interpreted cautiously and regularly updated as new data becomes available to ensure accurate and reliable insights for decision-making.

V. CONCLUSION

This work focused on developing a predictive model for energy consumption using XGBoost regression and SARIMAX. The model's performance metrics on both the training and test sets indicate its effectiveness in capturing the complex dynamics of energy systems. The results of the comparison between the XGBoost and SARIMAX models for energy consumption forecasting highlight the strengths and limitations of each approach. The XGBoost model demonstrates proficiency in capturing the cyclical patterns and general trends in energy consumption. However, it may struggle to accurately forecast peak consumption periods, often underestimating the magnitude of energy usage during these high-demand periods. On the other hand, the

SARIMAX model showcases its ability to capture both the cyclical patterns and accurately forecast peak consumption periods, providing reliable and accurate predictions. Therefore, the choice between the two models depends on the specific requirements of the forecasting task. The XGBoost model may be suitable for capturing general consumption patterns, while the SARIMAX model is more adept at capturing both cyclical trends and peak periods of energy usage.

The results of this research can serve as a valuable foundation for further improvements in energy consumption forecasting, enabling better resource planning and optimization. Future studies could explore incorporating additional variables and refining the model to enhance its accuracy in predicting peak consumption periods, thus aiding in effective energy management and decision-making.

REFERENCES

[1] Hadri, S., Najib, M., Bakhouya, M., Fakhri, Y., & El Arroussi, M. (2021). Performance Evaluation of Forecasting Strategies for Electricity Consumption in Buildings. Energies, 14(18), 5831.

[2] Bassi, A., Shenoy, A., Sharma, A., Sigurdson, H., Glossop, C., & Chan, J.H. (2021, June). Building energy consumption forecasting: A comparison of gradient boosting models. In The 12th International Conference on Advances in Information Technology (pp. 1-9).

[3] Shin, S.Y., & Woo, H.G. (2022). Energy Consumption Forecasting in Korea Using Machine Learning Algorithms. Energies, 15(13), 4880.

[4] Abbasi, R.A., Javaid, N., Ghuman, M.N.J., Khan, Z.A., & Ur Rehman, S. (2019). Short term load forecasting using XGBoost. In Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33 (pp. 1120-1131). Springer International Publishing.

[5] El Houda, B.N., Lakhdar, L., & Abdallah, M. (2022, October). Time Series Analysis of Household Electric Consumption with XGBoost Model. In 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-6). IEEE.

[6] Amitasree, P., Vamshi, G.R., & Devi, V.K. (2021, October). Electricity consumption forecasting using machine learning. In 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC) (pp. 1-8). IEEE.

[7] Elamin, N. and Fukushige, M., 2018. Modeling and forecasting hourly electricity demand by SARIMAX with interactions. Energy, 165,pp.257-268.

[8] Kaggle. (2022). Time Series Forecasting with Machine Learning - PJMW Hourly Dataset. Retrieved from https://www.kaggle.com/code/robikscube/time-series-forecasting-with-machine-learning-yt/input?select=PJMW_hourly.csv